The difficulty of diagnostics

Jesse Brunner

2020-10-15

Making black-and-white decisions in a grayscale world

The world we live in may have absolute, black and white truths, indeed we have built our social and judicial systems on this premise—an individual is guilty or innocent of a crime, needs public assistance or is trying to cheat the system, is merely exercising protected 1st or 2nd amendment rights or is threatening others' safety—but finding that clear line between the two is pretty damn tricky. We are almost always forced to use imperfect information to delineate these binary outcomes and no matter how black and white the reality may be, there are broad swaths of gray in the decision making.

Many years ago I spent an enlightening year as an AmeriCorp VISTA member helping administer a private energy assistance program. My job was to help people in need keep the power or heat on in their houses, mostly by helping pay down their overdue balances with money from private donations. I looked at individual applications, talked to them, and then decided whether we could help them and how. As a 22 year old college graduate, I had very little experience with people on the edge. I was often overwhelmed having to determine whether they were telling the truth, or whether they could follow through on their promises and payment plans I had arranged. It was gray as far as the eye could see. I was always a little relieved when people were clearly above the income cutoff; while I often felt bad that we couldn't help them, I wasn't forced to make a judgment about a person or their plight.

The problem was, I felt damned no matter what I did. If I was too liberal with assistance, I was surely helping people that were milking, if not cheating the system. Moreover, there was a real chance of not having enough money for the truly needy by the end of the season. On the other hand, if I were more skeptical or stringent, only helping where I felt certain that they were in over their head, through no (or little) fault of their own, and that they weren't coming to depend on our assistance, well then I was almost certainly excluding people that needed our help, but had a harder time making their case. I frequently sought the advice of the more seasoned women who ran other public programs, but learned I was going to quickly become cynical or apathetic if I kept trying to make these important decisions with very imperfect data. So I applied to graduate programs in biology in search of Truth[1], or at least unambiguous facts and rigorously tested hypotheses.

The irony is that that gray area of decision making is rampant in biology. Well, it's common in science in general, but maybe especially in my chosen field of disease ecology. The difference is, I think, that in the sciences we are encouraged to be open about the uncertainty, about the gray areas, and also that we have some

[1] With a capital "T".

useful tools to deal with it.

Is that an infection, or glowing lint?

Consider the most essential aspect of a study about disease, determining whether an individual is infected or not. The reality is that an individual falls into one category or the other, there are no “sort of” infected individuals[2], but we cannot know an individual’s state with complete certainty[3]. Instead, we measure something directly or indirectly related to the infection such as the worms in someone’s poo[4], the presence or quantity of pathogen DNA, or the degree to which antibodies in our blood react to a pathogen. Let us stick with antibodies.

If you’ve been infected with a pathogen and have recovered you have probably developed antibodies to many parts of that pathogen. So when someone draws your blood they can see if you have antibodies in it that react to those pathogen parts, called epitopes. If so, there’s some signal, usually a color change depending on how the test is put together. But you might have a little signal while I produce a big one. For instance, individuals that were recently infected, are immunosuppressed, or are just, in the words of a friend and colleague, “wonky,” and so they may be infected without producing many antibodies against the pathogen. Over all, there is a distribution of antibody reactions from infected individuals.

You might expect this will be easy and all of the samples from uninfected individuals would be very low, maybe even zero. However, you, me, a mouse, or a bird all produce antibodies that cross react with pathogens we have never seen. There is always a bit of a signal. So while there is a different distribution of antibody reactions from uninfected individuals, it might overlap a bit with the infected distribution. In other words, there is almost always a gray area (OK, purplish in this figure) in terms of antibody titers or concentrations that includes both uninfected and infected individuals.

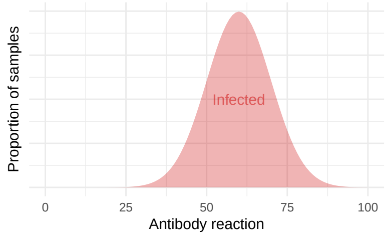

Figure 1: Hypothetical distributions of an antibody reaction (e.g., in an ELISA) from infected individuals.

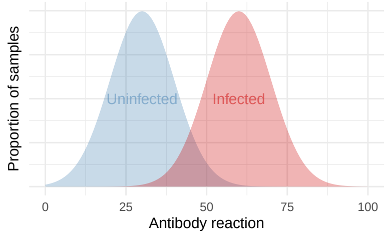

Figure 2: Hypothetical distributions of an antibody reaction (e.g., in an ELISA) from infected and uninfected individuals.

[2] Though there are large differences in the *intensity* of infections, disease symptoms, etc.

[3] At least not without experimentally infecting them!

[4] Yup! There’s a person who’s job is to look in those stool samples and count worms. It’s actually pretty neat!

Because we are trying to categorize individuals into infected and uninfected categories, we need to choose some threshold amount of the reaction. Above this threshold the test is scored as positive, below it, negative. However, unless there is perfect separation[5] (i.e., the two curves do not overlap) there will be some false positives (uninfected samples that test positive) or some false negatives (infected samples that test negative). While we can shift this threshold lower, say, to minimize the number of false negatives, by doing so we end up increasing the number of false positives! Or we could increase our threshold to ensure there were very few false positives, but that would come at the cost of more false negatives. There is an inherent trade-off between increased sensitivity—ensuring there are few or no false negatives—and specificity–ensuring there are few or no false positives.

The same is true of virtually every diagnostic method. One lab in which I worked determined whether black legged ticks were infected with the agent of Lyme disease using a reagent that made the bacterium fluoresce under a microscope, but so could pieces of lint that sometimes curled in just the right way to look like a bacterium, and of course if you didn't look at the right place or blinked at the wrong time, you might even miss this glowing signal. So we undoubtedly said that some uninfected ticks were infected (false positives) and some infected ticks were not (false negatives). There are sometimes ways to narrow this gray area—use a better diagnostic test, use a two-step test with different methods[6]—but the gray area is almost universally present[7] whether in diagnostic tests or in benefit determinations.

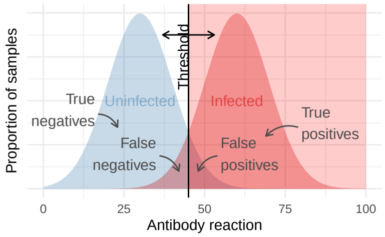

Figure 3: Hypothetical distributions of an antibody reaction (e.g., in an ELISA) from uninfected and infected individuals. Above some threshold amount of antibody reaction a sample is scored as positive (pink background) and below it is negative.

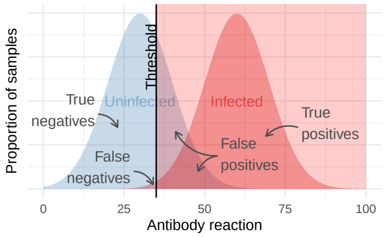

Figure 4: Same distributions, but the threshold has been moved to a lower, less stringent level so there are fewer false negatives, but more false positives.

Classification and ROC curves

So how do scientists deal with these gray areas? How do we decide where to draw the line between one category and another? Well, we acknowledge the trade-off between sensitivity and specificity and then find the threshold value that does the best job of achieving a balance between these competing interests given our goals.

One incredibly useful tool to think through these trade-offs is the ROC (Receiver operating characteristic) curve. This is the product of yet another detection problem, detecting enemy airplanes or ships using early versions of RADAR in World War II. The problem is the same as with detecting infection status, if not a bit more immediate. Is that blip I see on the screen a real plane or something else like a flock of birds or a phantom in the electronics? These curves help us see how good or bad our test might be and, perhaps even more importantly, better understand the trade-offs we face with imperfect tests.

Imagine we were to move the threshold in the previous figures from the far right (super stringent) to the far left (super lenient) and record the number or proportion of false negatives and false positives at each point. If we plotted 1 - false positive rate (aka sensitivity or the true positive rate) against the false positive rate (=1-specificity) we would have a nice curve sweeping an arc into the upper left corner. These curves describe the trade-off between sensitivity and specificity.

[5] Better diagnostic test have greater differentiation between the two distributions, but in most real-world cases, we cannot completely separate them.

[6] More on this later.

[7] Even physics has this problem. For instance, was that signal in the CERN accelerator really a new particle or just some noise from the detector or another particle or…? Their key advantage is that they can do the experiment a gazillion times until they are exceedingly confident that what they saw, and keep seeing, is real.

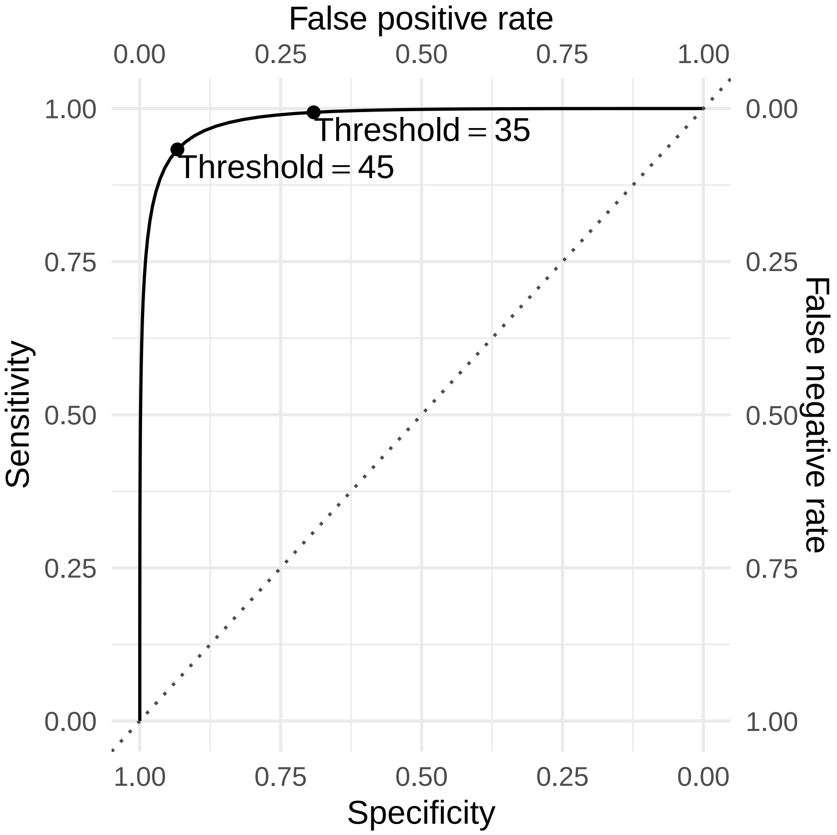

Figure 5: ROC plot for the hypothetical antibody test. The cutoffs from the prior figures are plotted along the curve. Notice the axis values are reverse from top to bottom and left to right. When a test is pefectly sensitive there are no false negatives.

Notice the two threshold we used above are plotted along this curve. The first was a threshold of 45, which resulted in a Sensitivity and Specificity of 0.933. That is, 93 out of 100 truly infected samples will test positive and 93 out of 100 truly uninfected samples will test negative; but seven in each case will be scored wrong. If we were to reduce the threshold to 35 we would increase the sensitivity to 0.994—over 99 out of 100 truly infected samples will test positive—but this comes at the cost of many more false positive tests, 31 out of 100 truly uninfected samples will test positive!

The beauty of these curves is that they help us visualize[8] how the rates of false negatives and false positives change as you move the threshold from more stringent (no false positives, but many false negatives) to less stringent (more false positives, but fewer false negatives). The ideal diagnostic test or radar system would produce a curve that is essentially a rectangle[9] where there is a threshold value that perfectly delineated airplanes from birds, infected from uninfected individuals, the needy from the conniving. In this case there would be no overlap between the red and blue distributions in Figs. 1 & 2.

However, in the real world there is almost always some noise. As the amount of noise increases (i.e., the red and blue distributions overlap more and more), the distinction between infected and uninfected, or plane and bird, becomes more and more ambiguous and the ROC curve approaches the one-to-one line[10] (dotted line in Fig. 3). This is essentially the same as flipping a coin and saying if it's heads then the person is infected. Or more directly, the test is useless. Most of the time our diagnostic test or radar system falls somewhere in between, which means we have a trade-off between certainty that we will not miss an airplane or infection (low chance of false negatives) and an increase in false alarms (false positives).

What is the right threshold? Is there a right threshold?

Where should a person draw the line, then? If you give the problem a bit of thought my guess is that you recommend finding the threshold that maximizes overall accuracy[11] or, equivalently, minimizes the total number of false positives and false negatives. More accurate[12] is better, right? Well yes... so long as the costs of false negatives and false positives are the same. If you were in charge of the radar system in WWII would you be more worried about the cost of scrambling your forces for no good reason in response to a false positive or about missing the first airplanes in an invasion due to overly stringent criteria and false negatives? Would you want the threshold that maximized accuracy or would you rather move that threshold a bit to ensure you had very, very few false negatives at the cost of perhaps a lot more false positives? The cost of scrambling your airplanes to intercept a phantom invader is not negligible, but is probably quite small relative to the cost of an enemy airplane going undetected in which case you might lose lives, a defensive post, or a battle. And so you will probably choose the less stringent threshold and a lot of false positives. In peacetime, however, your calculations probably change.

[8] Try this interactive app to get a better sense of how signal, noise, and thresholds interact.

[9] And the "area under the curve" (AUC), if you've run across this term, is one.

[10] An AUC of 0.5.

[11] Accuracy = $\frac{TP+TN}{\text{All tests}}$

[12] It is worth noting that if someone says a test if 97% accurate, you do not know if the inaccurate results are mostly false negatives or false positives or both. It is hard to summarize a test's performance in one single number!

The same is true when it comes to diagnostic tests. There is a debate raging over advice on who should get a mammogram and when. Epidemiologists have calculated that the combined costs of all of the false positive diagnoses—the unneeded treatments and surgery, not to mention the worry and expenses—outweigh the benefits of subjecting nearly all women over a certain age to the test in order to minimize the chances that a malignant tumor goes undetected. Or similarly, is it better to tweak the test to ensure we detect as many COVID-19 cases as possible, but inadvertently cause a bunch of uninfected people to have to quarantine themselves, or is it better to be able to trust that everyone who tests positive is really positive[13]? More on this later, but first we need to take a step back and see that the problem is actually much, much harder than we might like.

The signal gets swamped: Bayes rule and prevalence

Let us continue thinking about COVID-19 testing. There is surprisingly little publicly available information about diagnostic[14] sensitivity. In part this is because it is not clear how to determine who is *truly* infected and who is *truly* negative[15] and in part it is because the amount of virus (and antibodies) changes throughout an infection, varies among different classes of people, and so on. But let's simplify things and assume that real-time reverse-transcriptase PCR reactions, the basis of most tests, have a sensitivity of 0.95 (95%) and a specificity of 0.97 (97%). Not bad, right? (Most tests, especially those based on detecting viral antigens or antibodies, are almost certainly worse, especially on the sensitivity side of things, but let's just pretend.)

Now comes the wrinkle. Let's also assume that 1% of the population as a whole is infected. If we were to randomly[16] test 10,000 people we would expect 100 people to be really, truly infected and of those, 95 would be correctly identified, leaving 5 false negatives. I guess we could live with that, although if we were thinking of Ebola virus infections that might be worrying! Now what about the 9,900 uninfected people? The vast majority, 9,603, would be correctly told they are not infected, but 3% of them, or 297 uninfected people would be told they *are* infected. Or put another way, only 95/(95+297) = 24% of the people with positive test are actually infected! But wait, it can be even worse. In the earlier stages of the epidemic when only 0.1% of the population was infected, nine or ten of the truly infected individuals would be detected, but roughly 300 of the 9,990 *uninfected* people would get false positives, so only 10/(10+300) = ~3% of the positive tests would indicate true infections.

We can formalize the probability that a person is truly infected (I^+) given a positive test ($T+$) using Bayes rule[17].

$$\Pr(I^+|T^+) = \frac{\Pr(T^+|I^+) \times \Pr(I^+)}{\Pr(T^+)} \tag{1}$$

$$= \frac{\Pr(T^+|I^+) \times \Pr(I^+)}{\Pr(T^+|I^+) \times \Pr(I^+) + \Pr(T^+|I^-) \times \Pr(I^-)} \tag{2}$$

[13] One way to express the utility of a test is the positive predictive value (=TP/(TP+FP)), sometimes called precision, which basically tell us the probability that a positive test is actually a true positive, and, similarly, the negative predictive value (=TN/(TN+FN)), which is the reverse.

[14] As opposed to analytic sensitivity, which measures how little of the virus can be detected.

[15] Usually one compares their test to a "gold standard," which is assumed to be correct, but in this case with a novel disease how can one be sure? This is never an easy problem, but it's even harder in the middle of a fast-moving crisis!

[16] This is important in that if we just test sick people we would tend to have more infected people than the population average of 1%.

[17] This theorem is fundamental in statistics, but not widely appreciated or intuited by most of us without formal training. Even if the math escapes you, focus on the logic.

We can read this as, “The probability a person is infected, given that they test positive, is equal to the probability they test positive given that they are infected times the probability they really are infected, divided by the probability they test positive.” Does that help? No?

OK, let’s try this instead: The part that reads $\Pr(T^+|I^+)$, or the probability of testing positive given that you are positive? That is just sensitivity. The $\Pr(I^+)$ term is the probability[18] that you are actually infected, which is, all else equal, the prevalence in the population. The denominator in the fraction is then the probability of getting a positive test, whether a true positive $[\Pr(T^+|I^+) \times \Pr(I^+)]$ or a false positive $[\Pr(T^+|I^-) \times \Pr(I^-)]$. Note that false positives occur when a person is uninfected, which happens with probability $1 -$ prevalence, but are incorrectly scored as positive. If specificity is the probability of *correctly* scoring a negative as a negative, then $1 -$ specificity is the probability of *falsely* scoring a negative as a positive. So we could rewrite the equation like this:

$$\Pr(I^+|T^+) = \frac{\text{Sensitivity} \times \text{Prevalence}}{\Pr(\text{True positive}) + \Pr(\text{False positive})} \tag{3}$$

$$= \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity}) \times (1 - \text{Prevalence})} \tag{4}$$

Let us plug in the numbers from our example with 1% prevalence, a sensitivity of 0.95, and a specificity of 0.97:

$$\Pr(I^+|T^+) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + (1 - 0.97) \times (1 - 0.01)} \tag{5}$$

$$= 0.242, \tag{6}$$

which matches our previous calculation[19].

If we simply change the prevalence to 0.1% we get:

$$\Pr(I^+|T^+) = \frac{0.95 \times 0.001}{0.95 \times 0.001 + (1 - 0.97) \times (1 - 0.001)} \tag{7}$$

$$= 0.031, \tag{8}$$

which also matches our previous calculation.

The probability that a person is *actually* infected given a positive test changes a great deal with the prevalence of the infection in a population. Even with a very good diagnostic test, when prevalence is low a positive test still probably does not indicate a real infection! The true positives are simply swamped by all of the false positives. And the reverse is true when we consider the probability that a negative indicates a real negative (see the blue line in the figure).

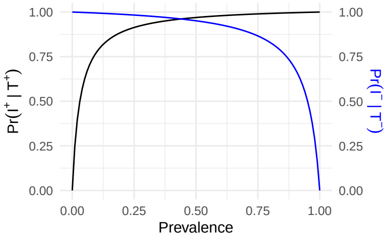

Figure 6: The probability a person is truly infected given a positive test (black line) or truly uninfected given a negative test (blue line) over a range of prevalence of infection in the population.

Before moving on, there is one important caveat to all of this. As noted above, we were assuming that people were tested at random, so the probability they were

[18] In Bayesian terms, the *prior* probability you were infected before you had a test result.

[19] Note that in a Bayesian sense we have updated our belief about whether this person is infected. We initially thought there was a 1% chance, based on the prevalence of the population, but after including new information, a positive test, we now think there is a 24% chance this person is infected.

infected before getting the test, the $\Pr(I^+)$ part, was just prevalence in the population. But if you go to the doctor because you are sick with flu-like symptoms and get tested, you are no longer a random person. The probability that you are infected is higher, probably much higher, than the general population. Maybe it's one in three instead of one in one-hundred. In that case, the probability you are actually infected given a positive test, like the one we described above, is 94%. So I am *not* recommending you ignore your doctor and the lab results she gives you! I am, however, urging caution about things like using antibody tests across a broad swath of the population to assign "immunity cards." Run the numbers and see how much faith you would have that someone testing positive for antibodies is actually immune[20].

What are we to do?

I hope I've convinced you that this seemingly simple task, sorting individuals into infected or uninfected bins, is actually much more difficult than it first seems. And there are always trade-offs lurking. Sure, we might gain sensitivity, but we do so at the cost of specificity. And depending on the context a positive test might not translate into a high probability that you are actually infected. So, what's a person to do?

First, it is worth thinking about what we are trying to accomplish with a test. Tests are used for all sorts of things and pretending they are used in only one way is to ignore all of the trade-offs and potential biases we just learned. If we are thinking about you in a doctor's office getting back test results (or conversely, you, the doctor, giving someone the news) it is worth thinking about how well the test performs, whether you had a good chance of having whatever the test was for in the first place, and probably looking for some confirmation from other independent tests and context. In short, the test is useful, but would you base your healthcare[21] on any one diagnostic test? Alternatively, if we are interested in trying to estimate the prevalence of infection in the population, we can correct our estimates so long as we know the sensitivity and specificity. We may not know whether *this* person or *that one* is infected, but we could have a pretty good idea that, say, 20% of the population was infected, on average^[This is part of the rationale for using fast, cheap, but crappy antibody-based tests for COVID-19. Yes, they suck, but they can still give us a window on what's happening and, importantly, we get the answers back quickly!.

Second, it is worth noting that we can be clever about our testing regime. For instance, we might employ a two-step approach for widespread testing. We would employ a test with high sensitivity, but low specificity, for instance and antibody test tune for maximum sensitivity. This would ensure that few truly infected people test negative, but would yield a *lot* of false positives. Those false positives would then be re-tested with a separate test with much higher specificity (e.g., a nucleic acid-based test) to sort out which of that big heap of "positives" from the first test were actually real. Most HIV screening is done this way and many have

[20] All of which is complicated by the facts that antibody titers vary in time and many tests measure non-neutralizing antobodies, as well as that infections can sometimes persist and hide out for long periods

[21] Or someone's access to government assistance or guilt or innocence.

recommended it for COVID-19 testing[22].

Third, it pays to think about the costs of false negatives *and* false positives[23]. What are the costs of false positives? Perhaps many, many people being quarantined for no good reason, not to mention all of the stress and worry. What are the costs of missing a COVID-19 infection? A person might ignore their other symptoms because they thought they were clear, perhaps even die. At a larger scale, the infected person who thinks they are COVID-free might go on to infect (many) others and so the infection spreads.

This highlights one other issue that crops up frequently when it comes to diagnostic testing: the right or optimal answer at an individual level is not necessarily the right or optimal answer for the population. A doctor might see the ghost of a tumor in a mammogram and recommend a biopsy if not a lumpectomy. Yes, it might have been a false positive, but the cost—an unnecessary procedure—seems small compared to letting a real tumor grow and, potentially, become much worse. An epidemiologist, who focuses on the whole population, might instead note that because tumors are fairly rare and mammograms imperfect, there are lots and lots of false positives, the costs of which collectively outweigh the benefits of catching a few tumors early. Both are "right," given their perspective, but they might strongly disagree about what is best. You could extend the same thinking to the TSA's security at airports or the Innocence Project's efforts to find and exonerate the wrongly convicted. Or even to all of those amazing people at offices across the nation trying to sort out who, among the many applicants, most need public assistance.

By focusing on just one type of cost (or one axis of the ROC curve) I think we miss the larger point that there are trade-offs and that the relative costs and benefits are, at least in principle, calculable. We may disagree on how to value different outcomes, but the underlying trade-off between false positives and false negatives is, in principle, objective. I think that being transparent about why we choose particular cutoffs, and what that means in terms of trade-offs, would be useful to our society more generally. At a minimum, though, I think that ROC curves, and the perspectives embedded in them, can help us have more productive discussions about where we should draw the lines.

[22] Also, I *think* that mortgage underwriters work this way…at least it would explain why they kept asking for the same information over and over again!

[23] My own opinion is that many conflicts boil down to some people focusing on one set of costs—such as missing a tumor or someone who honestly needs assistance—while others focus only on the costs of false positives—unnecessary health care interventions or welfare cheats.